A Most Parsimonious Reconstruction Problem on Phylogenetic Trees

AUTHOR(S):

Narushima, Hiroshi

CITATION:

Narushima, Hiroshi. A Most Parsimonious Reconstruction Problem on Phylogenetic Trees. 数理解析研究所講究録 1994, 871: 233-240

ISSUE DATE:

1994-05

URL:

http://hdl.handle.net/2433/84031

数理解析研究所講究録
第871巻 1994年 233-240

進化系統樹の最節約復元（MPR）問題について*
A Most Parsimonious Reconstruction Problem on Phylogenetic Trees

成嶋　弘（東海大・理・情報数理）
Hiroshi Narushima (Tokai Univ.)

abstract

A combinatorial optimization problem regarding assignments (called reconstructions) to a tree has been discussed in phylogenetic analysis. J. S. Farris, D. L. Swofford and W. P. Maddison have solved the problem of finding most parsimonious reconstructions on a completely bifurcating phylogenetic tree. We formulate mathematically the problem with its generalization to the case of any tree and call it the MPR problem. We present a solution for the generalized problem by introducing the concept of median interval obtained from sorting the endpoints of some closed intervals. The state set operation which plays an important role in the Farris–Swofford–Maddison method, is clarified by the concept of median interval. And then, with an explicit recursive formulation we generalize smoothly their method. Also, the computational complexity of our method is discussed. In the discussion, the PICK algorithm by Blum–Floyd–Pratt–Rivest–Tarjan is essential.

1. Introduction

The following optimization problem originated in cladistics (biological systematics and phylogenetics) has been proposed. Let $\mathbf{R}$ be the set of real numbers and $\mathbf{N}$ be the set of nonnegative integers. In particular, we use Ω to denote the set that may be either $\mathbf{R}$ or $\mathbf{N}$. Let $T = (V = V_O \cup V_H, E, \sigma)$ be any tree with the leaves evaluated by a weight function $\sigma : V_O \to \Omega$, where V is the set of nodes, V_O is the set of leaves, V_H is the set of internal nodes, and E is the set of branches. In phylogenetic trees, σ is called a *character state function*, each leaf is called an operational taxonomic unit, and each intermal node is called a hypothetical taxonomic unit. We call this tree an *el-tree*, where "el" is an abbreviation of "evaluated leaf". From an algorithmic point of view, we shall sometimes restrict Ω to $\mathbf{N}$. For an el-tree T, we define an assignment $\lambda : V \to \Omega$ such that $\lambda|V_O$ (the restriction of λ to V_O) $= \sigma$, that is, $\lambda(v) = \sigma(v)$ for each v in V_O, where $\lambda(v)$ is called a *state* of v under λ. This assignment is called a *reconstruction* on an el-tree T in phylogenetic analysis. For each branch e in E of an el-tree with a reconstruction λ, we define the *length* $l(e)$ of $e = \{u, v\}$ by $|\lambda(u) - \lambda(v)|$. Furthermore, for each reconstruction λ on an el-tree T, we define the *length* $L(\lambda)$ of λ by $L(\lambda) = \sum_{e \in E} l(e)$. Then $L^*(T)$ is defined by

$$L^*(T) = \min\{L(\lambda) | \lambda \text{ is a reconstruction on } T\}.$$

We here mention that $L^*(G)$ is well-defined. It is sufficient for us to consider the range

(*) This paper is a digest version of the reference [4]

of λ as the closed interval $[\min\sigma, \max\sigma]$ (written as Δ). Therefore, we can think of L as a function from the set $\{\lambda : V \to \Delta\}$ of reconstructions on T into Ω. When $\Omega = \mathbf{N}$, it is obvious that the minimum of L exists. When $\Omega = \mathbf{R}$, we see that the function L is continuous on the compact space, and so, the minimum of L exists. A *most parsimonious reconstruction* (MPR) on an el-tree T is a reconstruction λ such that $L(\lambda) = L^*(T)$. We denote the set of all MPRs on an el-tree T by $\mathsf{Rmp}(T)$.

The problem is as follows:

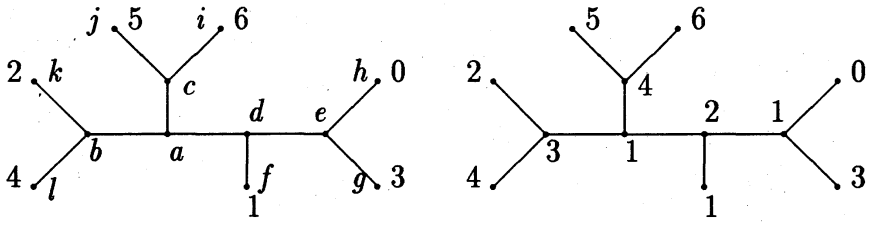
1. determine $L^*(T)$ for a given el-tree T,
2. find all MPRs on a given el-tree.

We call this problem the MPR problem. For the meaning of the MPR problem in cladistics the reader may refer to Swofford-Maddison [3] and Minaka [2].

In Fig. 1 we show an example for an el-tree T that is also given in [3] and an example for a reconstruction $\lambda : V \to \mathbf{N}$ on T. Then

$$\begin{aligned} L(\lambda) &= |\lambda(a) - \lambda(b)| + |\lambda(a) - \lambda(c)| + |\lambda(a) - \lambda(d)| + \cdots \\ &= 2 + 3 + 1 + \cdots = 16. \end{aligned}$$

We see later on that $L^*(T) = 10$. Throughout this paper, we use the el-tree T shown in Fig. 1 (i) whenever we illustrate results in this paper with an example.



(i) An el-tree T (ii) A reconstruction on T

Fig. 1.

2. Definitions

Let $\mathbf{P}$ be the set of positive integers. A closed interval $\{x | a \leq x \leq b\}$ in Ω is denoted by $[a, b]$. We denote the closed interval $[1, n]$ in $\mathbf{N}$ by $[n]$, that is, $[n] = \{1, 2, \ldots, n\}$. Let I and J be any closed intervals in Ω. We denote the "nearest" distance between I and J by $d(I, J)$, that is,

$$d(I, J) = \min_{x \in I, y \in J} |x - y|.$$

Particularly, the "nearest" distance $d(x, I)$ between a real number x and a closed interval I is

$$d(x, I) = \min_{y \in I} |x - y| = d([x, x], I).$$

Note that $d(I,J) = 0$ does not necessarily mean $I = J$ and also the triangle inequality $d(I,K) \leq d(I,J) + d(J,K)$ does not necessarily hold. Therefore, this "distance" d between closed intervals of Ω is not a distance function. For any family I_i $(i \in [m])$ of closed intervals in Ω we define a function $D : \Omega \to \Omega$ by

$$D(x) = \sum_{i \in [m]} d(x, I_i).$$

We might denote $D(x)$ by $D(x, I_1, I_2, \cdots, I_m)$ or $D(x, I_i : i \in [m])$ to avoid ambiguity. The minimum of $D(x)$ is denoted by $D^{\min}(I_1, I_2, \cdots, I_m)$ or $D^{\min}(I_i : i \in [m])$. Let $I_i = [a_i, b_i]$ $(i \in [m])$ be any family of closed intervals in Ω. Let all the endpoints a_i and b_i of I_i $(i \in [m])$ be sorted in ascending order and then be arranged as follows:

$$x_1 \leq x_2 \leq \cdots \leq x_m \leq x_{m+1} \leq \cdots \leq x_{2m}.$$

Then we call the closed interval $[x_m, x_{m+1}]$ in Ω the *median interval* of the closed intervals I_i $(i \in [m])$, which is the key concept in this paper and denoted by $\mathrm{med}\langle I_1, I_2, \cdots, I_m \rangle$ or $\mathrm{med}\langle I_i : i \in [m] \rangle$.

Lemma 1. *Let $I_i = [a_i, b_i]$ $(i \in [m])$ be any closed intervals in Ω and $x_i \leq x_{i+1}$ $(i \in [2m-1])$ be the sorted sequence of the endpoints of I_i $(i \in [m])$ in ascending order. Then we have*

$D(x, I_i : i \in [m]) = D^{\min}(I_i : i \in [m])$ *if and only if* $x \in \mathrm{med}\langle I_i : i \in [m] \rangle$.

Let $T = (V, E)$ be a rooted (directed) tree, where V is the set of nodes and E $(\subseteq V \times V)$ is the set of branches. For each u and v in V, we write $u \to v$ when $(u, v) \in E$, that is, u is a *parent* of v (or v is a *child* of u). For each u and v in V, u is called an *ancestor* of v (or v is called a *descendent* of u), written $u \Rightarrow v$, if there is a sequence of nodes $u = u_1, u_2, \cdots, u_n = v$ in V such that $u_i \to u_{i+1} (i \in [n-1])$, which is called a *path* in T. We call a leaf (a node without a child) of a rooted tree a *sink* to avoid ambiguity. For each u in V, we denote a *subtree* of T induced from a subset $\{v \in V | u \Rightarrow v\}$ (including u) of V by $T_u = (V_u, E_u)$. Note that u is the root of T_u.

Let $T = (V_O \cup V_H, E, \sigma)$ be an el-tree rooted at r in $V = V_O \cup V_H$. The rooted el-tree is sometimes written as $T^{(r)}$ to show the root r explicitly. In addition, if r is a leaf, i.e., $r \in V_O$ and s is its unique child, we represent the rooted tree as (T_s, r) to vizualize the structure. In this case, the subtree T_s is called the *body* of the tree T; otherwise, i.e., if the root is not a leaf, the body of T is T itself.

For each node u in the body of a rooted el-tree T, we assign a closed interval $I(u)$ of Ω recursively as follows.

$$I(u) = \begin{cases} [\sigma(u), \sigma(u)] & \text{if } u \text{ is a sink,} \\ \mathrm{med}\langle I(v) : u \to v \rangle & \text{otherwise.} \end{cases}$$

We call $I(u)$ the *characteristic interval* of a node u and so I is called the *characteristic interval map* on T.

Let T be a completely bifurcating el-tree rooted at a node r. Then we see that $I(u)$ is just the *Farris interval* of a node u in p.204of [3], and that $I(r)$ is just the *MPR-set* S_r of a

node r in p.212 of [3], which is the set of states that may be assigned to node r in an MPR. It is shown later on that $I(r)$ is the MPR-set of a node r in an el-tree $T^{(r)}$. These facts show that the concept of characteristic interval, the essence of which is a median interval, is a unified generalization of the two concepts, Farris interval and MPR-set.

Let T be again an el-tree rooted at a node r in V. Then, we define a number $l^*(u)$ of Ω recursively for each node u of the body of T as follows.

$$l^*(u) = \begin{cases} 0 & \text{if } u \text{ is a sink,} \\ \sum_{u \to v} l^*(v) + D^{\min}(I(v) : u \to v) & \text{otherwise.} \end{cases}$$

It is shown later on that $l^*(r) = L^*(T)$ which is defined in Introduction. So, we call l^* the *minimum length map* on T.

We here give examples for computing $I(u)$ and $l^*(u)$ for each u in V. Let T be an el-tree shown in Fig. 1 (i) and $T^{(a)}$ be the tree T rooted at node a (Fig. 2 (i)). Then we obtain I (Fig. 2 (ii)) and l^* (Fig. 2 (iii)).

For example,

$$I(a) = \text{med}\langle [2,4], [5,6], [1,1] \rangle = [2,4]$$

since the endpoints 2, 4, 5, 6, 1, 1 are sorted as 1, 1, 2, 4, 5, 6, and

$$\begin{aligned} l^*(a) &= l^*(b) + l^*(c) + l^*(d) + D^{\min}([2,4],[5,6],[1,1]) \\ &= 2 + 1 + 3 + 4 = 10 \end{aligned}$$

since

$$\begin{aligned} D^{\min}([2,4],[5,6],[1,1]) &= d(2,[2,4]) + d(2,[5,6]) + d(2,[1,1]) \\ &= 0 + 3 + 1 = 4. \end{aligned}$$

Also, $T^{(f)} = (T_d, f)$ and I, l^* (with respect to $T^{(f)}$) are illustrated in Fig. 2 (iv) – (vi).

3. Theorems

Let T be a rooted el-tree (T_s, r) and I be the characteristic interval map on T. Then we define recursively a reconstruction λ (with $\lambda(r) = \sigma(r)$) on T as follows:

(i) $\lambda(s) \in \text{med}\langle [\lambda(r), \lambda(r)], I(t) : s \to t \rangle$,

(ii) for all v such that $u \to v$,
$\lambda(v) \in \text{med}\langle [\lambda(u), \lambda(u)], I(w) : v \to w \rangle$,

where σ is a character state function of T. In general, when we define a function $f : X \to Y$, for a subset B of Y, "$f(x) \in B$ " means that "B is the set of elements which may be assigned to x". Note that the above definition of λ is defined in the direction from the root to sinks. We here write $\mathsf{Rmp2}(r, s)$ for the set of all reconstructions constructed by the above definition. Let $\lambda_{<u>}$ denote the restriction $\lambda | V_u$ of a reconstruction λ on T to a subtree T_u of T. Then the set $\mathsf{Rmp2}(r, s)$ is also defined recursively as follows: $\lambda_{<s>} \in \mathsf{Rmp2}(r, s)$ if and only if (1) $\lambda(s) \in \text{med}\langle [\lambda(r), \lambda(r)], I(t) : s \to t \rangle$ and (2) for all t such that $s \to t$,

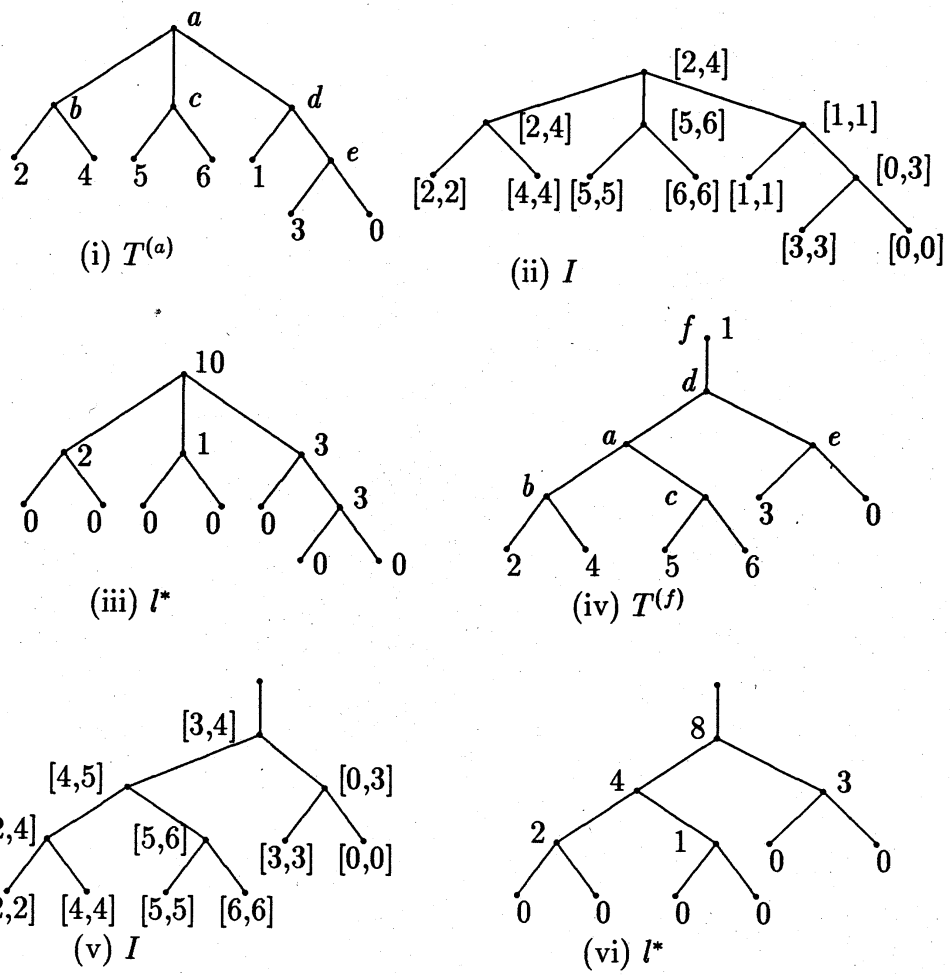

a
b
c
d
e
2
4
5
6
1
3
0
(i) $T^{(a)}$
[2,4]
[2,4]
[5,6]
[1,1]
[2,2]
[4,4]
[5,5]
[6,6]
[1,1]
[0,3]
[3,3]
[0,0]
(ii) I
10
2
1
3
3
0
0
0
0
0
0
0
(iii) l^*
f
1
d
a
e
b
c
3
0
2
4
5
6
(iv) $T^{(f)}$
[3,4]
[4,5]
[0,3]
[2,4]
[5,6]
[3,3]
[0,0]
[2,2]
[4,4]
[5,5]
[6,6]
(v) I
8
4
3
2
1
0
0
0
0
0
0
(vi) l^*

Fig. 2.

$\lambda_{<t>} \in \mathsf{Rmp2}(s,t)$. Note that $\lambda_{<s>}$ (with $\lambda(r) = \sigma(r)$) can be considered a reconstruction on T.

Let T be a rooted el-tree $T^{(r)}$ with the root r in V_H. Let I be the characteristic interval map on T. Then we define recursively a set $\mathsf{Rmp1}(r)$ of reconstructions on T as follows: $\lambda \in \mathsf{Rmp1}(r)$ if and only if (1) $\lambda(r) \in I(r)$ and (2) for each s such that $r \rightarrow s$, $\lambda_{<s>} \in \mathsf{Rmp2}(r,s)$.

Theorem 1. *Let T be an el-tree. Then we have*

(i) $L^*(T) = l^*(r)$ *and* $\mathsf{Rmp}(T) = \mathsf{Rmp1}(r)$ *when* $T = T^{(r)}$ $(r \in V_H)$,

(ii) $L^*(T) = l^*(s) + d(\sigma(r), I(s))$ *and* $\mathsf{Rmp}(T) = \mathsf{Rmp2}(r,s)$ *when* $T = (T_s, r)$.

The following corollary, a generalization of Theorem 2 in [3], is obtained from the definition of $\mathsf{Rmp1}(r)$ and Theorem 1 (i).

Corollary 1. *Let I be the characteristic interval map on a rooted el-tree $T^{(r)}$ $(r \in V_H)$. Then $I(r)$ is the MPR-set (written as S_r) of a node r, which is the set of states that may be assigned to r in an MPR.*

We now give an example for generating MPRs on an el-tree T. From Theorem 1 and the recursive definitions of Rmp1(r) or Rmp2(r,s), we see that the enumeration method is a two-pass algorithm which consists of the first pass: the determination of the characteristic interval map I on T defined recursively in the direction from the sinks to the root and the second pass: the determination of each element of $\mathsf{Rmp}(T)$ defined recursively in the direction from the root to sinks. Note that the choice of v in Step (ii) (or the choice of t in Step (2)) of the definition of $\mathsf{Rmp2}(r,s)$ may be carried out by the depth first search or the breadth first search. Note further that the essential part in both of the two passes is the computation of median intervals. Let T be an el-tree shown in Fig. 1 (i) and $T^{(a)}$ be the tree T rooted at node a (Fig. 2 (i)). Then we have the map I on $T^{(a)}$ (Fig. 2 (ii)) and by using the depth first search on the set V of nodes, each MPR λ on T is defined and shown in Table 1:

$\lambda(a) \in [2,4]$
  $\boldsymbol{\lambda(a) = 2}$
    $\lambda(b) \in \mathrm{med}\langle[2,2],[2,2],[4,4]\rangle = [2,2]$
      $\boldsymbol{\lambda(b) = 2}$
    $\lambda(c) \in \mathrm{med}\langle[2,2],[5,5],[6,6]\rangle = [5,5]$
      $\boldsymbol{\lambda(c) = 5}$
    $\lambda(d) \in \mathrm{med}\langle[2,2],[1,1],[0,3]\rangle = [1,2]$
      $\boldsymbol{\lambda(d) = 1}$
        $\lambda(e) \in \mathrm{med}\langle[1,1],[3,3],[0,0]\rangle = [1,1]$
          $\boldsymbol{\lambda(e) = 1}$
      $\boldsymbol{\lambda(d) = 2}$
        $\lambda(e) \in \mathrm{med}\langle[2,2],[3,3],[0,0]\rangle = [2,2]$
          $\boldsymbol{\lambda(e) = 2}$
  $\boldsymbol{\lambda(a) = 3}$ *and* $\boldsymbol{\lambda(a) = 4}$ (These cases are omitted)

$\lambda \backslash u$	a	b	c	d	e	f	g	h	i	j	k	l
λ_1	2	2	5	1	1	1	3	0	6	5	2	4
λ_2	2	2	5	2	2	1	3	0	6	5	2	4
λ_3	3	3	5	1	1	1	3	0	6	5	2	4
λ_4	3	3	5	2	2	1	3	0	6	5	2	4
λ_5	3	3	5	3	3	1	3	0	6	5	2	4
λ_6	4	4	5	1	1	1	3	0	6	5	2	4
λ_7	4	4	5	2	2	1	3	0	6	5	2	4
λ_8	4	4	5	3	3	1	3	0	6	5	2	4

Table 1: $\mathsf{Rmp}(T) = \mathsf{Rmp1}(a)$

4. Computational Complexities

One sees in the previous sections that the description of the problem and the algorithms is simple but the proof of validity of the algorithms is not so simple. The complexity analysis is also not so difficult, because all the key concepts are recursively defined.

First of all, considering the well-definability of $L^*(T)$ for a given el-tree T, which is mentioned in Introduction, we see that it is sufficient for solving the MPR problem to examine δ^h reconstructions on T, where $\Delta = [\min \sigma, \max \sigma]$, δ is the cardinality of Δ and $h(= |V_H|)$ is the number of internal nodes of T. When $\Omega = \mathbf{N}$, we could solve the MPR problem by the primitive finite algorithm, i.e., the method of checking all posibilities, since $\delta^h < \infty$, but the complexity order is exponential. When $\Omega = \mathbf{R}$, note that δ^h is not finite.

We now discuss about the algorithmic complexity of our algorithms for the following four problems:

1. determine $L^*(T)$ for a given el-tree T,
2. find any one MPR on a given el-tree,
3. enumerate all MPRs on a given el-tree.

We must appreciate that the description of the "sorted" sequence of end points of closed intervals in Ω is often used in the previous sections, but the number of comparisons required to "select" the i-th smallest of n numbers (denoted by $f(i,n)$) is essential in the complexity analysis of our algorithms. Therefore, our time complexity analysis is based on the following result (Theorem 1 in p.450 of [1]) for the selection algorithm called PICK by Blum et al. [1].

PICK Theorem. *The number $f(i,n)$ of comparisons required to select the i-th smallest of n numbers is at most a linear function of n, i.e., $f(i,n) = O(n)$.*

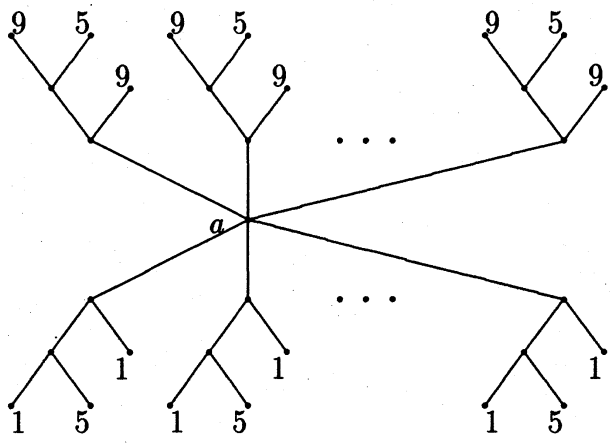
Theorem 2. *The complexity of our algorithm for Problem 1 is $O(n)$ for the number n of nodes in a given el-tree.*

Theorem 3. *The complexity of our algorithm for Problem 2 is $O(n)$ for the number n of nodes in a given el-tree.*

When $\Omega = \mathbf{R}$, we do not discuss the complexity of algorithm for Problem 3 by the obvious reason.

Proposition 1. *When $\Omega = \mathbf{N}$, there is an el-tree T such that the number of all MPRs on T is exponential for the number n of the nodes.*

Proof. Consider the rooted el-tree $T^{(a)}$ shown in Fig. 3. □



A tree with $10m + 1$ modes and at least 5^{2m} MPRs.

Fig. 3.

References

[1] M. Blum, R. W. Floyd, V. Pratt, R. L. Rivest, and R. E. Tarjan, Time bounds for selection, Journal of Computer and System Sciences 7 (1973) 448 - 461.

[2] N. Minaka, Parsimony, phylogeny and discrete mathematics: combinatorial problems in phylogenetic systematics (in Japanese: with English summary), Natural History Research, Chiba Prefectural Museum and Institute, Vol.2 No.2 (1993) 83 - 98.

[3] D. L. Swofford and W. P. Maddison, Reconstructing ancestral character states under Wagner parsimony, Mathematical Biosciences 87 (1987) 199-229.

[4] M. Hanazawa, H. Narushima and N. Minaka, Generating most parsimonious reconstructions on a tree: a generalization of the Farris-Swofford-Maddison method, to appear.

[5] M. Hanazawa and H. Narushima, A more efficient algorithm for MPR problems on an el-tree, to appear.

[6] H. Narushima and N. Misheva, On the positions of Acctran and Deltran in the MPR-poset, to appear.